Jurnal Sains dan Teknologi
Volume 4 No. 3 | Juni 2023 |pp: 51-56
E-ISSN : 2714-8661 | DOI : https://doi.org/10.55338/saintek.v5i1.1093

Peningkatan Kinerja Random Forest Melalui Seleksi Fitur Secara Pca Untuk Mendeteksi Penyakit Diabetes Tahap Awal

Zenny Susanti[1*], Pahala Sirait[2], Erwin Setiawan Panjaitan[3]

[1,2,3]Program Studi Magister Teknologi Informasi Universitas Mikroskil Medan, Indonesia
Email Penulis Korespondensi: [1]194212091@students.mikroskil.ac.id

Abstrak− Diabetes merupakan penyakit yang banyak diderita oleh semua kalangan usia, salah satu penyakit yang menyiksa dengan angka kematian yang tinggi, maka perlu dilakukan diagnosis diabetes stadium awal karena penyakit ini jarang terdeteksi secara dini oleh tubuh yang terkena diabetes. Dalam proses diagnosis, masalah yang paling sering terjadi adalah waktu pengambilan keputusan dan ketepatan diagnosis yang ditemui dalam proses klasifikasi. Atribut penting dalam pengambilan keputusan diabetes tahap awal, sehingga perlu diketahui atribut utama diabetes. Seringkali diperoleh hasil yang berbeda dalam proses diagnostik karena banyak atribut yang digunakan dalam pengambilan keputusan. Sehingga perlu dilakukan proses pengurangan atribut diabetes. Metode Principal Component Analysis (PCA) dapat digunakan untuk mereduksi data dengan dimensi besar dan mereduksi peringkat atribut. Proses klasifikasi dapat dilakukan dengan menggunakan algoritma random forest dan mendapatkan tingkat akurasi dalam proses klasifikasi. Pengujian dengan perbandingan 90:10, 80:20, 70:30 hasil prediksi penyakit menggunakan metode random forest memiliki hasil dan tingkat akurasi yang berbeda.

Kata Kunci: Data Mining, Diabetes, PCA , Random Forest, RapidMiner

Abstract− Diabetes is a disease that affects many people of all ages, one of the torturous diseases with a high mortality rate, it is necessary to make a diagnosis for early-stage diabetes because this disease is rarely detected early by the body affected by diabetes. In the diagnostic process, the most frequent problems are the decision-making time and the accuracy of the diagnoses encountered in the classification process. Attributes are important in decision-making for early-stage diabetes, so it is necessary to know the main attributes of diabetes. Often different results are obtained in the diagnostic process because many attributes are used in decision making. So it is necessary to do a process of reducing the attributes of diabetes. Principal Component Analysis (PCA) method can be used for data reduction with large dimensions and the ranking of attributes to be reduced. The classification process can be done using the random forest algorithm and get a level of accuracy in the classification process. Tests with a ratio of 90:10, 80:20, 70:30 the results of disease prediction using the random forest method have different results and accuracy levels.

Keywords: Data Mining, Diabetes, PCA, Random Forest, RapidMiner

I. PENDAHULUAN

Seiring dengan perkembangan teknologi, kebutuhan akan informasi secara cepat dan tepat menjadi hal yang sangat penting saat ini dalam mengambil keputusan. Pentingnya melakukan prediksi terhadap sebuah kasus akan memberikan beberapa peluang dalam memperbaiki situasi dan kondisi ataupun kemungkinan-kemungkinan kejadian yang tidak diinginkan terjadi dimasa yang akan mendatang. Hal ini sudah banyak diterapkan dalam segala jens bidang ilmu seperti ilmu managemen perekonomian, ilmu dalam bidang bisnis, ilmu geologi, geografi, pertanian, hingga ilmu dalam dunia medis. Salah satunya dalam dunia medis sangat bermanfaat bagi banyak kalangan, karena penerapan ini sangat banyak digunakan terhadap penyakit-penyakit serius sehingga dapat memperlambat bahkan menghilangkan penyakit yang telah didiagnosa sebelumnya [1].

Kinerja machine learning dalam perkembangan teknologi informasi sangat membantu dalam dunia medis, dimana kalangan dokter dan tim medis lainnya mendapatkan kesempatan dalam meningkatkan kualitas kerja tim medis. Salah satu jenis penyakit berbahaya yang dapat diprediksi merupakan penyakit diabetes dimana penyakit ini merupakan jenis penyakit yang banyak menyebabkan kelumpuhan, kerusakan yang mengerikan pada penderita lanjutan hingga kematian. Hal tersebut terjadi karena keterlambatan diagnosis. Sehingga pentingnya dilakukan diagnosis untuk membantu proses medis yang lebih komplek lagi dan merupakan proses yang penting dalam sebuah peneltian dibidang medis. Permasalahan yang sangat sering dihadapi dan sering terjadi dalam hal diagnosis waktu yang sangat lama dalam proses prediksi terhadap gejala dan kurangnya tingkat akurasi dalam proses klasifikasi [2].

Dalam proses penerapan sesuatu maka dibutuhkan machine learning dan machine learning tidak dapat beroperasi tanpa algoritma, seleksi fitur merupakan salah satu komponen yang sangat penting dalam menentukan akurasi dalam proses menjalankan algoritma klasifikasi sehingga pentingnya mengetahui atribut utama pada sebuah penyakit. Sering didapatkan hasil yang berbeda ketika mendiagnosis penyakit diabetes awal, sehingga dipilih fitur atribut penting yang didasarkan pada efisiensi diagnosis [3]. Keputusan diagnosis penyakit yang berbeda merupakan salah satu tantangan yang paling penting untuk analisis data diabetes tahap awal. Umumnya keputusan diagnosis didasarkan kepada pemeriksaan kepada pasien dan juga pengalaman dokter. Pada pemeriksaan yang dilakukan tidak semua kondisi atribut terpenuhi, tetapi tetap harus dilakukan proses diagnosis [4].

Jurnal Sains dan Teknologi
Volume 4 No. 3 | Juni 2023 |pp: 51-56
E-ISSN : 2714-8661 | DOI : https://doi.org/10.55338/saintek.v5i1.1093

Kesalahan pada proses diagnosis akan mempengaruhi hasil pada pasien ataupun dokter, jadi diagnosis merupakan tugas kompleks yang membutuhkan keahlian tinggi dan pengalaman. Sistem komputer dapat membantu para dokter sebagai alat memprediksi dan diagnosis penyakit diabetes, penelitian medis selalu berususan dengan data yang besar, penanganan data yang besar dengan tidak benar dapat mempengaruhi keakuratan terhadap hasil. Dataset yang digunakan merupakan hasil dari informasi-informasi yang disediakan oleh situs yang akurat pada bidangnya [5]. Pemilihan fitur yang tepat dapat mempersingkat waktu dan menghemat biaya dalam melakukan diagnosis salah satu proses pemilihan fitur dapat dilakukan reduksi atribut yang digunakan untuk mengurangi data dan menghapus atribut yang tidak relevan, berlebihan, dalam sebuah dataset. Pengurangan atribut telah menjadi langkah penting dalam pengenalan pola dan tugas Machine Learning [6]. Hal ini menginspirasi penulis sehingga menambahkan proses reduksi atribut di dalam pengambilan keputusan pada proses klasifikasi dalam meningkatkan diagnosis penyakit diabetes tahap awal.

Seleksi fitur dilakukan untuk mengidentifikasi subset atribut yang optimal [7]. Seleksi fitur dapat mengurangi dimensi data dan waktu pelatihan [8]. [9] Menjelaskan bahwa mengikutsertakan seleksi fitur dapat memberikan dua keuntungan yaitu dapat membantu mengurangi curse of dimensionality dan mengkomputasikan fitur yang penting dapat membantu interpretasi data. Contoh metode untuk seleksi fitur seperti, Principal Componen Analysis (PCA), Chi-square, Wrapper based feature selection dan filter. dimana berinteraksi dengan pengklasifikasi untuk mengevaluasi kegunaan fitur sehingga menghasilkan subset fitur yang lebih baik. Metode ini lebih lambat dibandingkan dengan metode filter based feature selection, Neural Network [10] [11]. Principal component analysis (PCA) adalah teknik untuk mengurangi dimensi atribut dari dataset, meningkatkan interpretabilitas tetapi pada saat yang sama meminimalkan hilangnya informasi, banyak penelitian dan buku yang telah ditulis dan juga bahkan ada buku tentang varian PCA untuk tipe data tertentu [12]. PCA dapat digabungkan dengan metode lain yang akurasinya dikenal keakurasiannya cukup tinggi, dimana PCA berfungsi untuk melakukan reduksi dimensionalitas dan dengan pemilihan fitur pada PCA yang hanya memilih nilai eigen yang paling besar maka ketika digabungkan dengan metode lain akan menaikan tingkat akurasi klasifikasi dan terutama sekali mempercepat hasil pengenalannya [13]

Pada penemuan tingkat akurasi dari proses klasifikasi terhadap penyakit diabetes, penelitian terdahulu melakukan pengujian terhadap diabetes yang memiliki banyak perkembangan variasi sehingga diilakukan ekstraksi terhadap pemiihan data set meggunakan metode beberapa metode klasifikasi secara bersamaan seperti Decision tree, Naïve Bayes, K-Nearest Neighbor memiliki tingkat akurasi terhadap ekstraksi fitur data set sebesar 75,65% [14]. Dan pada penelitian lainnya, memperlihatkan metode decition tree dalam menghasilkan nilai akurasi yang lebih tinggi terhadap pemilihan fitur dengan menggeneralisasi fitur yang optimal dari dataset, sehingga dapat meningkatkan akurasi klasifikasi dengan pencapaian 98,00% [15]. Pernyataan beberapa penelitian terdahulu menyebutkan dalam proses klasifikasi menggunakan random forest lebih memiliki tingkat akurasi yang tinggi daripada algoritma klasifikasi naïve bayes, decision tree, K-nearest neighbor (Xaverius et al., 2020).

Metode Radom forest telah membuktikan keberhasilannya dalam prediksi penyakit diabetes tahap awal dengan melakukan perbandingan tiga algoritma klasifikasi machine learning yaitu Suport Vector Machine, Naive Bayes dan Random Forest digunakan dalam percobaan ini untuk mendeteksi diabetes secara dini. Hasil penelitian terlihat masalah regresi dan klasifikasi data dan merupakan salah satu algoritma machine learning terbaik adalah algoritma random forest yang digunakan di berbagai bidang dengan pencapaian dengan tingkat akurasi 97,88 %, dan tool yang digunakan adalah WEKA [17].

II. METODE PENELITIAN

Tahapan Penelitian

Pada penyusunan tesis ini penulis melakukan beberapa tahapan dalam penerapan metodologi untuk menyelesaikan masalah. Berikut merupakan langkah-langkah penyelesaian ini adalah:

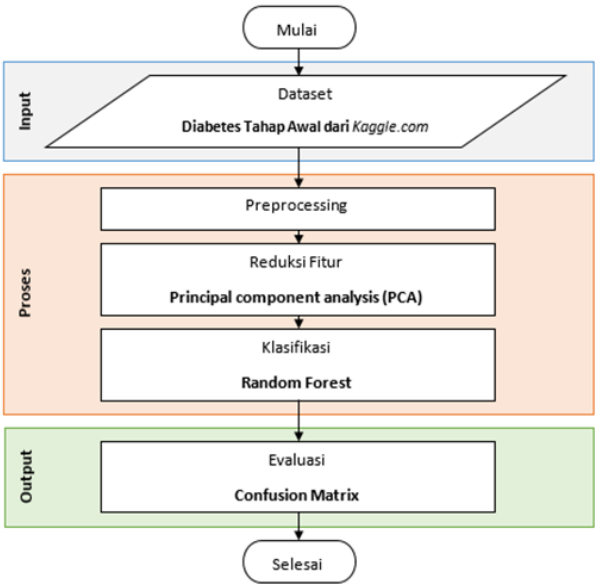
1. Studi Literatur
   Dalam tahapan ini dilakukan pengumpulan data sebagai bahan referensi yang berhubungan dengan penelitian yang dilakukan, seperti: penyakit Diabetes Tahapan Awal, dataset penyakit diabetes, algoritma Random forest dan Principal Component Analysis (PCA)
2. Analisis Masalah
   Pada tahap ini dilakukan proses untuk mengidentifikasi data yang dibutuhkan, masalah dan tantangan yang harus diselesaikan dan menjelaskan solusi yang diusulkan untuk menyelesaikan masalah dan tantangan yang ada.
3. Preprosesing data
   Pada tahapan preprosesing data ini digunakan untuk melakukan penyesuaian data terhadap algoritma ataupun metode yang akan digunakan, pada penelitian tahap *preprocessing* akan dilakukan dalam dua cara yaitu *cleaning* dan *binning* menggunakan IBM IBM SPSS Modeler 18.0, sehingga hasilnya dapat diproses untuk pengujian selanjutnya.
4. Tahap Pengujian

Jurnal Sains dan Teknologi
Volume 4 No. 3 | Juni 2023 |pp: 51-56
E-ISSN : 2714-8661 | DOI : https://doi.org/10.55338/saintek.v5i1.1093

Melakukan pengujian terhadap seleksi fitur penyakit diabetes tahap awal dengan menggunakan metode PCA. Kemudian menerapkan algoritma *Random Forest* dalam proses prediksi dengan dan tanpa seleksi fitur menggunakan RapidMiner 9.8.1. hasil sesuai pada tahapan analisa dengan melihat nilai dari *confusion matrix*.

5. Evaluasi
   Pada tahap ini, dilakukan evaluasi terhadap hasil pengujian yang telah dilakukan untuk mengambil kesimpulan dan
   saran.

Adapun tahapan pada penelitian ini adalah sebagai berikut.



Gambar 1. Metode Penelitian

Metode PCA (Principal Component Analysis)

Principal Component Analysis (PCA) melakukan pemetaan/transformasi set data dari dimensi lama ke dimensi baru dengan memanfaatkan teknik dalam aljabar linear, tanpa memerlukan parameter tertentu dalam memberikan keluaran hasil pemetaanya. PCA memerlukan masukan data yang mempunyai sifat zero-mean pada setiap fiturnya. Sifat zero-mean pada setiap fitur data bisa didapatkan dengan mengurangkan semua nilai dengan rata-ratanya. Set data X dangan dimensi MxN, dimana M adalah jumlah data dan N adalah jumlah fitur akan tampak seperti berikut

$$X = \begin{bmatrix} X_{11} & X_{12} & X_{1j} & X_{1N} \\ X_{21} & \dots & \dots & X_{2N} \\ X_{31} & \dots & \dots & X_{3N} \\ X_{i1} & \dots & \dots & X_{iN} \\ X_{M1} & X_{M2} & X_{Mj} & X_{MN} \end{bmatrix}$$

Untuk fitur ke-j, semua nilai pada kolom tersebut dikurangi dengan rata-ratanya, diformulasikan dengan:

$$X_{ij}=X_{ij} - \bar{X}_j \quad (1)$$

i = 1,2 …, M, dan j adalah kolom ke-j

Selanjutnya dilakukan perhitungan matriks kovarian dari matriks X, yaitu C_X. Formula yang digunakan adalah dot-product pada setiap fitur.

$$C_X=\frac{1}{M}X^T.X \quad (2)$$

N adalah jumlah fitur, sedangkan X^T adalah matriks transpos dari X.

$$C_X = \frac{1}{M} \begin{bmatrix} X_{1\,1} & X_{2\,1} & X_{i\,1} & X_{M\,1} \\ X_{1\,2} & \dots & \dots & X_{M\,2} \\ X_{1i} & \dots & \dots & X_{Mi} \\ X_{1\,N} & X_{2\,N} & X_{i\,N} & X_{N\,M} \end{bmatrix} X \begin{bmatrix} X_{11} & X_{12} & X_{1j} & X_{1N} \\ X_{21} & \dots & \dots & X_{2N} \\ X_{i1} & \dots & \dots & X_{iN} \\ X_{M\,1} & X_{M\,2} & X_{M\,j} & X_{M\,N} \end{bmatrix}$$

$$= \frac{1}{M} \begin{bmatrix} X_{11} & X_{12} & X_{1j} & X_{1N} \\ X_{21} & \dots & \dots & X_{2N} \\ X_{i1} & \dots & \dots & X_{iN} \\ X_{N1} & X_{N2} & X_{Nj} & X_{NN} \end{bmatrix}$$

Pada matriks C_X, elemen ke-ij adalah inner-product antara baris matriks X_T dengan kolom matriks X. Jika Y adalah matriks set data hasil pemetaan dan C_Y adalah matriks kovarian dari Y, yang diharapkan dalam PCA adalah:

1. Semua elemen selain diagonal utama dama C_Y harus nol. Maka, C_Y harus matriks diagonal. Dengan kata lain, Y adalah matriks terdekorelasi.
2. Peletakan dimensi dalam Y dari kiri ke kanan diurutkan menurun.

$$Xu = \lambda u \quad (3)$$

Dengan mencari matriks ortonormal P dimana Y = PX dan $C_Y = \frac{1}{M}YY^T$ adalah matriks diagonal, dan kolom dari P adalah komponen utama (principal components) dari X, persamaan C_Y bisa dijabarkan sebagai berikut:

$$C_Y = PC_XP^T \quad (4)$$

Metode Random Forest

Pada algoritma *random forest* ini juga merupakan algoritma yang terdiri dari algoritma *decision tree* yang dikenal sebagai pohon prediksi dimana setiap pohon keputusan bergantung kepada nilai random vector dan di jadikan sebagai sampel secara bebas pada semua pohon (*tree*) dalam forest tersebut. Adapun rumus penyelesaian masalah yang digunakan dalam algoritma *random forest* ini sebagai berikut:

$$l(y) = argmax_c(\sum_{n=1} I_{h_n(y)=c}) \quad (5)$$

Algoritma ini akan lebih mudah dan cepat dianalisa jika menggunakan apikasi pendukung yaitu Rapidminer

Jurnal Sains dan Teknologi
Volume 4 No. 3 | Juni 2023 |pp: 51-56
E-ISSN : 2714-8661 | DOI : https://doi.org/10.55338/saintek.v5i1.1093

yang membantu mempermudah memproses semua akses data yang tersedia.

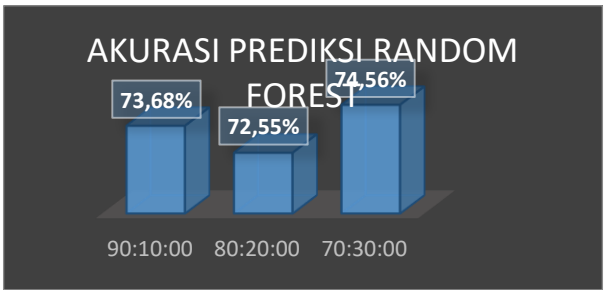
III. HASIL DAN PEMBAHASAN

Pada bagian ini dilakukan proses prediksi dan peningkatan akurasi klasifikasi, terlebih dahulu dilakukan klasifikasi pada dataset dengan menggunakan algoritma *Random Forest*, kemudian dilakuakan reduksi dataset dengan menggunakan metode Principal Component Analysis (PCA). Pada reduksi dataset menggunakan metode Principal Component Analysis (PCA) reduksi atribut berdasarkan dengan kombinasi atribut dengan nilai *Cumulative Varience* atau nilai ranking yang terbesar. Setelah dilakukan reduksi dan mendapatkan kombinasi atribut dengan nilai Eigenvalue terbesar kembali melakukan tahapan klasifikasi untuk melihat hasil dan juga akurasi dari proses klasifikasi. Tujuan reduksi atribut untuk mengetahui atribut serta tingkat akurasi yang paling optimal pada dataset setelah dilakukan proses reduksi atribut dan kombinasi atribut. Setelah dilakukan proses klasifikasi pada dataset yang tidak menglami reduksi atribut, kamudian mengukur tingkat akurasi dari dataset berdasarkan dengan algoritma yang digunakan. Seluruh pengujian dilakukan dengan menggu

Adapun hasil keseluruhan dari pengujian menggunakan *random forest* tanpa dilakukan reduksi sebagai berikut ini:

Tabel 1. Keterangan Jumlah Data *Training* dan Data Uji

No	Jumlah Dataset	Rasio Data	Jumlah Data *Training*	Jumlah Data Uji	Akurasi prediksi Random Forest
1	762	90:10	686	76	73,68%
2	762	80:20	610	152	72,55%
3	762	70:30	533	229	74,56%



Gambar 2. Grafik Akurasi *Random Forest* Tanpa PCA

Hasil Penggunaan PCA

Kemudian setelah dilakukan klasifikasi dan mengukur tingkat akurasi pada dataset yang tidak terdapat reduksi atribut, selanjutnya melakukan reduksi 1 atribut sampai dengan 4 atribut berdasarkan nilai *Cumulative Varience* untuk melakukan perangkingan kombinasi atribut, rangking terbesar merupakan data yang harus dilakukan seleksi atribut yang direduksi. Adapun hasil metode Principal Component Analysis (PCA) dapat dilihat di bawah ini:

Tabel 2. Keterangan Hasil Uji Metode PCA

Component	Standart Deviation	Proportion Of Variance	Cumulative Variance
Pregnancies	1,438	0,259	**0,259**
Glucose	1,318	0,127	**0,476**
BloodPressure	1,005	0,126	**0,602**
SkinThickness	0,954	0,144	**0,715**
Insulin	0,882	0,097	**0,813**
BMI	0,829	0,086	**0,899**
DiabetesPedigreeFunction	0,648	0,052	**0,951**
Age	0,625	0,049	**1**

Pada tabel 2 didapatkan perolehan hasil nilai masing-masing komponen atau atribut mulai dari nilai terendah hingga nilai tertinggi dan proses reduksi atribut berdasarkan dengan nilai *Cumulative Varience* yang paling kecil hingga terbesar juga.

Pengujian Metode *Random Forest* Dengan PCA

Pada tahapan ini dilakukan pengujian metode random forest dengan menggunakan reduksi atribut sebanyak 4 (empat) kali untuk mengetahui akurasi nilai yang lebih baik lagi, reduksi atribut sebelumya menggunakan metode PCA (*Principal Component Atribute*), adapun nilai tertinggi berada pada data *age, Diabetes Pedigree Function, BMI dan Insulin* menyebabkan terbentuknya dataset baru yang jauh lebih komplek, klasifikasi yang dilakukan mengunakan metode *random forest* setelah dilakukan reduksi atribut memiliki nilai akurasi yang berbeda-beda.

Pada proses pengujian tetap dilakukan perbandingan rasio untuk membagi data latih dan data uji yang digunakan pada tools *RapidMiner* versi 9.8.1, tetap meggunakan rasio 90:10, 80:20 dan 70:30 pada dataset yang tersedia dimana jumlah data terbesar sebagai data *training* atau data latih dan selebihnya bagian kecil menjadi data Uji. Tujuan dari pembagian rasio untuk mengetahui tingkat kualitas dan akurasi dataset yang tersedia, setiap rasio dilakukan pada tahapan reduksi 1 atribut, reduksi 2 atribut, reduksi 3 atribut hingga reduksi 4 atribut, digunakan 4 kali reduksi atribut sesuai pada jumlah nilai tertinggi hasil pengujian reduksi dataset menggunakan PCA (*Principal Component Atribute*) atau sebagian dari jumlah atribut

Jurnal Sains dan Teknologi
Volume 4 No. 3 | Juni 2023 |pp: 51-56
E-ISSN : 2714-8661 | DOI : https://doi.org/10.55338/saintek.v5i1.1093

yang tertera. Adapun beberapa hasil pengujian tersebut sebagai berikut ini:

Tabel 3. Perbandingan Akurasi Prediksi

No	Reduksi	Rasio	Tingkat Akurasi	
			Random Forest - PCA	Random Forest
1	Reduksi 1 Atribut	90:10	75,00%	90:10 = 73,68% 80:20 = 72,55% 70:30 = 74,56%
		80:20	73,20%	
		70:30	74,12%	
2	Reduksi 2 Atribut	90:10	**77,63%**	
		80:20	73,86%	
		70:30	75,88%	
3	Reduksi 3 Atribut	90:10	75%	
		80:20	73,20%	
		70:30	74,12%	
4	Reduksi 4 Atribut	90:10	76,32%	
		80:20	74,51%	
		70:30	74,56%	

IV. KESIMPULAN

Berdasarkan dari penjelasan permasalahan dan uraian setiap BAB serta hasil dari penelitian, maka dapat ditarik kesimpulan: Proses klasifikasi menggunakan algoritma *Random Forest* pada dataset menghasilkan tingkat akurasi yang berbeda – beda, sebelum dilakukan reduksi atribut nilai tertinggi sebesar 74,56% pada rasio dataset 70:30 dan nilai terendah sebesar 72,55% pada rasio 80:20 sementara pada setiap dilakukan reduksi atribut dengan tingkat akurasi yang paling rendah pada reduksi 1 atribut rasio 80:20 dengan memiliki akurasi sebesar 73,20% dan paling tinggi pada reduksi 2 atribut dengan akurasi sebesar 77,63%, hal ini dikarenakan yang keputusan mengalami reduksi, Proses klasifikasi dalam meningkatkan kinerja Random forest maka dilakukan reduksi atribut sebanyak 4 (empat) kali dengan kombinasi atribut dengan nilai *Cumulative Varience* atau nilai ranking yang terbesar untuk mengetahui akurasi nilai yang lebih baik lagi,. adapun nilai tertinggi berada pada data *age, Diabetes Pedigree Function, BMI dan Insulin*, Hasilnya adalah Metode Principal Componen Analysis (PCA) dapat digunakan untuk menyeleksi atribut dimana memiliki nilai kombinasi tertinggi sebagai atribut yang direduksi dan meningkatkan *performance* klasifikasi *Random Forest.*

V. REFERENCE

[1] J. Ma, "Machine Learning in Predicting Diabetes in the Early Stage," *Proc. - 2020 2nd Int. Conf. Mach. Learn. Big Data Bus. Intell. MLBDBI 2020*, pp. 167–172, 2020, doi: 10.1109/MLBDBI51377.2020.00037.

[2] Z. Lv and L. Qiao, "Analysis of healthcare big data," *Futur. Gener. Comput. Syst.*, vol. 109, pp. 103–110, 2020, doi: 10.1016/j.future.2020.03.039.

[3] U. California, I. Irvine, and A. Serikat, "Machine Learning dalam Memprediksi Diabetes Sejak Dini," pp. 167–172, 2021.

[4] G. Hendro, T. B. Adji, and N. A. Setiawan, "Penggunaan Metodologi Analisa Komponen Utama (PCA) untuk Mereduksi Faktor-Faktor yang Mempengaruhi Penyakit Jantung Koroner," Semin. Nas. ScrETec, pp. 1–5, 2012.

[5] A. S. Albahri et al., "Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review," J. Med. Syst., vol. 44, no. 7, 2020, doi: 10.1007/s10916-020-01582-x.

[6] M. A. C, "Reduksi Atribut Pada Dataset Penyakit Jantung dan Klasifikasi," vol. 4, pp. 994–1006, 2020, doi: 10.30865/mib.v4i4.2355.

[7] I. S. Thaseen, C. A. Kumar, and A. Ahmad, "Integrated Intrusion Detection Model Using Chi-Square Feature Selection and Ensemble of Classifiers," Arab. J. Sci. Eng., vol. 44, no. 4, 2019, doi: 10.1007/s13369-018-3507-5.

[8] R. AlShboul, F. Thabtah, N. Abdelhamid, and M. Al-diabat, "A visualization cybersecurity method based on features' dissimilarity," Comput. Secur., vol. 77, 2018, doi: 10.1016/j.cose.2018.04.007.

[9] A. Thakkar and R. Lohiya, "Attack classification using feature selection techniques: a comparative study," J. Ambient Intell. Humaniz. Comput., vol. 12, no. 1, 2021, doi: 10.1007/s12652-020-02167-9.

[10] S. Islam Ayon and M. Milon Islam, "Diabetes Prediction: A Deep Learning Approach," Int. J. Inf. Eng. Electron. Bus., vol. 11, no. 2, pp. 21–27, 2019, doi: 10.5815/ijieeb.2019.02.03.

[11] V. R. Balasaraswathi, M. Sugumaran, and Y. Hamid, "Feature selection techniques for intrusion detection using non-bio-inspired and bio-inspired optimization algorithms," J. Commun. Inf. Networks, vol. 2, no. 4, 2017, doi: 10.1007/s41650-017-0033-7.

[12] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," Informatics Med. Unlocked, vol. 19, 2020, doi: 10.1016/j.imu.2020.100330.

[13] C. Lei et al., "A comparison of random forest and support vector machine approaches to predict coal spontaneous combustion in gob," Fuel, vol. 239, no. September 2018, pp. 297–311, 2019, doi: 10.1016/j.fuel.2018.11.006.

[14] A. Azrar, M. Awais, Y. Ali, and K. Zaheer, "Data mining models comparison for diabetes prediction," Int. J. Adv. Comput. Sci. Appl., vol. 9,

no. 8, pp. 320–323, 2018, doi: 10.14569/ijacsa.2018.090841.

[15] N. Sneha and T. Gangil, "Analisis diabetes mellitus untuk prediksi awal menggunakan pemilihan fitur yang optimal," vol. 0, 2019.

[16] F. Xaverius, H. Jayawardana, and S. H. Hidayat, "Comparing euclidean distance and nearest neighbor algorithm in an expert system for diagnosis of diabetes mellitus," Enferm. Clin., vol. 30, pp. 374–377, 2020, doi: 10.1016/j.enfcli.2019.07.121.

[17] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," Sistemasi, vol. 10, no. 1, p. 163, 2021, doi: 10.32520/stmsi.v10i1.1129.

[18] F. M. Putra and I. S. Sitanggang, "Classification model of air quality in Jakarta using decision tree algorithm based on air pollutant standard index," in IOP Conference Series: Earth and Environmental Science, 2020, vol. 528, no. 1. doi: 10.1088/1755-1315/528/1/012053.

[19] P. Appiahene, Y. M. Missah, and U. Najim, "Predicting Bank Operational Efficiency Using Machine Learning Algorithm: Comparative Study of Decision Tree, Random Forest, and Neural Networks," Adv. Fuzzy Syst., vol. 2020, 2020, doi: 10.1155/2020/8581202.

[20] L. Benali, G. Notton, A. Fouilloy, C. Voyant, and R. Dizene, "Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components," Renew. Energy, vol. 132, pp. 871–884, 2019, doi: 10.1016/j.renene.2018.08.044.

[21] R. Adennia, "IMPLEMENTASI CONVOLUTIONAL NEURAL NETWORK UNTUK EKSTRAKSI FITUR CITRA DAUN DALAM KASUS DETEKSI PENYAKIT PADA TANAMAN MANGGA MENGGUNAKAN RANDOM FOREST - UMM Institutional Repository," 2021. https://eprints.umm.ac.id/79344/ (accessed Apr. 03, 2022)